Discovery of Novel Metabolic Pathways in PGDBs

Luciana Ferrer
Alexander Shearer
Peter D. Karp

Bioinformatics Research Group
SRI International
Introduction

- We propose a computational method for the discovery of functional gene groups from annotated genomes
- The method can potentially be used for finding:
  - Novel pathways
  - Protein complexes or other kinds of functional groups
  - Genes that are functionally related to a starting gene of interest
- The method relies on sequence information only
- For now, restricted to prokaryotes
**Method Overview**

- Target Genome
  - Pairwise gene functional similarity score computation
  - Scores for target gene pairs
  - > thr
  - Candidate finder
  - genes in target genome

- Reference Genomes
  - Group functional similarity score computation
  - Compilation of known info
  - Report

SRI International Bioinformatics
Method Overview

1. Pairwise functional similarity scores: For all pairs of genes in the target genome find a measure of the probability that the genes are functionally related.

2. Candidate finder: Find all cliques (set of nodes linked to all others) in a network where
   - nodes are genes and,
   - edges are given when the above scores are above a threshold.

3. Group functional similarity scores: For each candidate group find a measure of the functional relatedness of its members. Optionally filter out groups with low score.

4. Generate Report: For each candidate group gather all available information to facilitate analysis.
Pairwise functional similarity scores

- Estimated using Genome Context (GC) methods
  - Use assumptions about the evolutionary processes to find associations between genes that might point to functional interactions
  - Uses the set of reference genomes to infer interactions (currently 623 bacterial genomes from BioCyc version 14.5)
  - Methods: Phylogenetic profiles, Gene neighbor, Gene fusion, Gene cluster
- Currently using only Gene Neighbor method, which is by far the best performing of the four
**Phylogenetic Profile Method**

- **Assumption:** *Genes whose products function together tend to evolve in a correlated fashion*
  - they tend to be preserved or eliminated together in a new species
- For each gene in the target genome create a binary vector with
  - a 1 in component i if the gene has a homolog in genome i
  - a 0 otherwise

**Score:** similarity between these vectors

![Diagram of Genomes and Genes](image.png)
Gene Neighbor Method
(Bowers 2004)

- **Assumption:** Genes whose products function together tend to appear nearby, at least in some genomes

- **For each gene pair**
  - Find the location of the *best* homologs of both genes in each of the reference genomes
  - For genomes that contain homologs of both genes, compute the relative distance between them
  - Score: a p-value for the observed distances
Results of Genome Context Methods

- Results on *E. coli K12*
  - Positive examples are gene-pairs in the same metabolic or signaling pathway or the same protein complex
  - All other pairs of genes of known-function are negative examples

- At this operating point:
  - 6869 pairs are labeled as positives
  - Around 28% of the positives are found
  - Only 0.1% of the negative samples are labeled as positives
  - But, this percent corresponds to 5044 negatives
Group Functional Similarity Scores

- For each candidate group find the reference genomes G that are enriched for the genes in the group.
- A genome G will be enriched for the group if:
  - A large fraction of the genes in the group have homologs in G, and
  - A small fraction of all the genes in the target genome have homologs in G.
List of genes with all known info about each
List of organisms enriched for group
List of organisms depleted for group
Phylogenetic similarity with known pathways from Metacyc
  - As phylogenetic profile method for genes but now for gene groups
  - Create binary vectors with a 1 if the organism is enriched for the candidate group
  - For each Metacyc pathway or complex, create a binary vector with a 1 for organisms that contain it
  - Compare these vectors with the one for the candidate
Report

- Genome context scores between gene pairs in the group
- BLAST E-values between gene pairs in the group
- Known pathways or complexes involving at least two genes from the group
- Genome context information
  - For each gene, list the relative position in all the organisms for which it has a homolog
Performance on *E. coli* K-12

- EcoCyc version 14.5 contains 944 protein complexes and 340 pathways curated from the literature
  - Of which 103 complexes and 175 pathways contain more than four genes
- Decide a candidate is correct if at least 70% of its genes are in a known pathway or protein complex
- We declare a pathway or complex as found by our method if at least 70% of its genes are included in some candidate
- Only consider candidates and pathways/complexes with more than 4 genes
  - Algorithm is less reliable for smaller groups
  - For candidates of size 2, it’s only as reliable as the genome neighbor method alone
### Results at Different Operating Conditions

<table>
<thead>
<tr>
<th>Percent of edges in network</th>
<th>Minimum number of enriched orgs</th>
<th>Number of candidates</th>
<th>Percent of correct candidates</th>
<th>Number of pathways found</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15%</td>
<td>0</td>
<td>1130</td>
<td>13%</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>312</td>
<td>19%</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>155</td>
<td>25%</td>
<td>42</td>
</tr>
<tr>
<td>0.07%</td>
<td>0</td>
<td>413</td>
<td>22%</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>150</td>
<td>29%</td>
<td>38</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>86</td>
<td>35%</td>
<td>13</td>
</tr>
</tbody>
</table>

- The percent of edges in the “actual” network for E. coli is 0.07%
- The predicted 0.07% contains some of those edges, but also many false positives
- So, you might want to include more edges to catch more of the positives
# Example 1: Rediscovered Pathways

Some examples of *E. coli K-12* pathways or complexes that are found by the proposed method.

<table>
<thead>
<tr>
<th>Pathway or Complex</th>
<th># genes in pathway or complex</th>
<th># matching genes in candidate</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td>Histidine biosynthesis</td>
<td>8</td>
<td>8</td>
<td>Perfect match</td>
</tr>
<tr>
<td>Tryptophan biosynthesis</td>
<td>5</td>
<td>5</td>
<td>Perfect match</td>
</tr>
<tr>
<td>ATP synthase</td>
<td>8</td>
<td>8</td>
<td>Perfect match</td>
</tr>
<tr>
<td>NADH:ubiquinone oxidoreductase I</td>
<td>13</td>
<td>13</td>
<td>Five additional genes: hycE/D/F and hyfH/G</td>
</tr>
<tr>
<td>Flavin biosynthesis I</td>
<td>6</td>
<td>5</td>
<td>One missing gene: ribF</td>
</tr>
</tbody>
</table>
Example 2: Nascent Biosynthetic Pathway

<table>
<thead>
<tr>
<th>Gene</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>moaA</td>
<td>b0781 molybdopterin biosynthesis protein A</td>
</tr>
<tr>
<td>moaB</td>
<td>b0782 molybdopterin biosynthesis protein B</td>
</tr>
<tr>
<td>moaC</td>
<td>b0783 molybdopterin biosynthesis protein C</td>
</tr>
<tr>
<td>moaE</td>
<td>b0785 molybdopterin synthase large subunit</td>
</tr>
</tbody>
</table>

- Missed getting moaD by very little (a slightly lower score on the pairwise functional similarity scores would have allowed us to find it)
- This a known biosynthetic pathway, but the exact pathway has not been elucidated yet and, hence, does not exist in EcoCyc
- This is one case that would count as an error in our statistics though it is really not an error
### Example 3

<table>
<thead>
<tr>
<th>Gene</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>dacA</td>
<td>b0632 D-alanyl-D-alanine carboxypeptidase, fraction A; penicillin-binding protein 5</td>
</tr>
<tr>
<td>dacC</td>
<td>b0839 penicillin-binding protein 6</td>
</tr>
<tr>
<td>dacD</td>
<td>b2010 DD-carboxypeptidase, penicillin-binding protein 6b</td>
</tr>
<tr>
<td>lipA</td>
<td>b0628 lipoate synthase monomer</td>
</tr>
<tr>
<td>rlpA</td>
<td>b0633 rare lipoprotein RlpA</td>
</tr>
</tbody>
</table>

- A RlpA-RFP fusion accumulates at cell division sites
- dacACD involved in peptidoglycan biosynthesis and cell morphology
### Example 4

<table>
<thead>
<tr>
<th>Gene</th>
<th>Product</th>
</tr>
</thead>
<tbody>
<tr>
<td>rsxE b1632</td>
<td>integral membrane protein of SoxR-reducing complex</td>
</tr>
<tr>
<td>rsxG b1631</td>
<td>member of SoxR-reducing complex</td>
</tr>
<tr>
<td>rsxD b1630</td>
<td>integral membrane protein of SoxR-reducing complex</td>
</tr>
<tr>
<td>rsxB b1628</td>
<td>member of SoxR-reducing complex</td>
</tr>
<tr>
<td>nth b1633</td>
<td>endonuclease III; specific for apurinic and/or apyrimidinic sites</td>
</tr>
</tbody>
</table>

- rsxABCDE predicted to form a membrane-associated complex
- Involved in regulation of soxS which participates in removal of superoxide and nitric oxide and protection from organic solvents
- nth has been shown to act in the process of base-excision DNA repair
Future Work

Two main obvious directions

- Instead of using a single genome context method, use them all in combination
  - Not trivial, we need training data (a gold standard) to find the combination function
  - Have an initial solution that is about to get into the system
- Relax the condition of the candidates being cliques in the network
  - Maybe some genes in the pathways are only related to some percent of the other genes in the pathway
Candidates for E. coli K-12

- Reports for the *E. coli K-12* candidates available in: http://brg.ai.sri.com/pwy-discovery/ecoli.html
- More documentation on the method and reports available from that page
- Better Web interfaces will be available in the future
- Applicable to other organisms
- Contact us if you are interested in more information pkarp@ai.sri.com, lferrer@ai.sri.com
References

- Paper on genome context methods:
  http://www.biomedcentral.com/1471-2105/11/493

- A paper on the pathway discovery method is under preparation